Task 1: Root Cause Analysis

After analyzing the logs from the RAG-based internal chatbot system, I identified two critical problems affecting performance and user experience:

Problem 1: Incorrect or Outdated Answers from PDFs

Hypothesis:

The system frequently retrieves document chunks from Archived Design Docs (PDFs) that contain obsolete or irrelevant information. This leads to user dissatisfaction, especially when the questions involve time-sensitive topics such as policies, revenue data, or technical benchmarks.

Component Involved: Retrieval

Data Source: Archived Design Docs (PDFs)

Evidence:

28% of all feedback involving PDF chunks resulted in a thumbs-down — the worst performance of all sources.

Example queries like “What is our company’s policy on GPT-4?” and “What are the current performance benchmarks?” pulled content from PDFs that was outdated or versioned incorrectly.

In contrast, the Engineering Wiki had an 85% thumbs-up rate, suggesting that live sources produce more accurate answers.

Conclusion:

The retrieval system appears biased toward selecting dense but outdated PDF content. This lowers the chatbot’s factual accuracy and hurts trust.

Problem 2: Latency Violates SLA

Hypothesis:

The Generation component, powered by LLaMA 3 (70B), is responsible for slow responses, especially when queries retrieve long or multiple chunks. This increases the number of tokens sent to the generator, slowing the model's processing time.

Component Involved: Generation

Data Source: Mixed (mainly long or multi-chunk inputs from various sources)

Evidence:

23 out of 81 queries (about 28.4%) exceeded the SLA latency limit of 3,500ms.

Many slow queries asked for summaries or company policies, which tend to trigger longer input contexts and output lengths.

Some responses took over 5 seconds, strongly indicating that generation time — not retrieval — is the primary bottleneck.

Conclusion:

Latency violations are primarily caused by the LLaMA 3 generator being overloaded with verbose input. Optimization is needed to reduce token load or adjust how input context is constructed.